Probing the Heavy Flavor Content in $t\bar{t}$ Events and Using $t\bar{t}$ Events as a Calibration Tool at CMS

Roberta Volpe on the behalf of CMS collaboration
Universita' degi Studi and INFN Perugia, via A. Pascoli 1, 06123 Perugia, Italy

We present two analyses dedicated to measure the ratio of branching ratios of the top quark, $R = B(t \to Wb)/(t \to Wq)$ (where $q = d, s, b$), using ttbar events with either one or two prompt isolated leptons (e or mu) in the final state. Furthermore the framework of the dileptonic analysis was used also for a feasibility study of the measurement of b-tagging efficiency, by assuming the R value to be the Standard Model one. Data-driven techniques to control the background in the selected events are discussed and the expected simulation results are presented.

1. Introduction

Top quarks decay mostly to Wb, while the final states Wd and Ws are suppressed by the square of the CKM matrix elements $|V_{td}|$ and $|V_{ts}|$. Besides single top studies, $|V_{tb}|$ can be obtained also through top pairs production, by measuring $R = B(t \to Wb)/(t \to Wq)$, with $q = d, s, b$, and assuming that exactly 3 generations of quarks exist, as the Standard Model (SM) predicts; indeed, by imposing the unitarity of the 3×3 CKM matrix, such ratio is $R = |V_{tb}|^2/(|V_{td}|^2 + |V_{ts}|^2 + |V_{tb}|^2) = |V_{tb}|^2$. Without any assumption on the number of generations of quarks, an R measurement is still useful to put constraints on V_{tb} and, more importantly, it can give a clue on the existence of a fourth generation; indeed in such scenario, R is appreciably less than the SM value [1]. The most recent R measurement obtained by CDF with ~ 162 pb^{-1} is $R > 0.61$ at 95 % C.L. [2]; DØ measured R simultaneously with the $t\bar{t}$ cross section and obtained the value $R = 0.97^{+0.09}_{-0.-08}$ and a limit $R > 0.79$ at 95 % C.L. with ~ 900 pb^{-1} [3]. The direct measurement of the CKM element $|V_{tb}|$ (predicted by the SM as $|V_{tb}| = 0.999133^{+0.000044}_{-0.000043}$) [4]) is possible only by means of the study of single top production and currently the only available measurements are from DØ [5] and CDF [6] experiments. In the CMS experiment [7], two feasibility studies of the R measurement have been carried on, one using selected semileptonic $t\bar{t}$ events [8] and described in Sec. 3, the other using selected dileptonic $t\bar{t}$ events [9] as described in Sec. 4. Both the analysis use data-driven methods in order to estimate the irreducible background contribution and consider the number of b-tagged jets as the physical observable, therefore the b-tagging efficiency must be fixed to a value obtained from an independent measurement. Furthermore, the framework of the analysis can be used also to the aim to perform a measurement of b-tagging efficiency by assuming the R value to be the SM one; such study was performed for the dilepton channel and is described in Sec. 4.3.2.

2. General Method

The parameter $R = B(t \to Wb)/B(t \to Wq)$ is measured by counting the number of jets originating from b-quark (b-jets) in $t\bar{t}$ events. The number of b-tagged jets depends, beyond the R value itself, on the b-tagging efficiency (ϵ_b) and the mis-tagging probability (ϵ_q). Therefore, the probability to have a given number i of b-tagged jets is a function of R, ϵ_b and ϵ_q. It is called $\varepsilon_i(R;\epsilon_b,\epsilon_q)$ and can be expressed as follows:

$$\begin{aligned}\varepsilon_i(R;\epsilon_b,\epsilon_q) &= R^2 P_i(t\bar{t} \to bWbW)\\ &+2R(1-R)P_i(t\bar{t} \to bWqW)\\ &+(1-R)^2 P_i(t\bar{t} \to qWqW)\end{aligned} \tag{1}$$

where q can represent an s or d quark and each P_i (probability for a definite $t\bar{t}$ decay of having i b-tagged jets in the final state) depends on ϵ_b and ϵ_q. This function is used to fit the distribution of the number of b-tagged jets (n_{btag}) to measure the value of the R parameter. In order to identify the flavor of the jets, specific algorithms are used. For this study, the *Track Counting* (TC) and *Jet Probability* (JP) [10] algorithms are used to tag the b-jets. The efficiency of the TC and JP algorithms can be measured in QCD events with reconstructed jets containing muons. The p_{Trel} method [11] exploits the distribution of the relative transverse momentum of the muon with respect to the jet to estimate the number of b jets present in data.

3. Semi-leptonic $t\bar{t}$ analysis

The final state of the semi-leptonic $t\bar{t}$ decay channel (one $W \to q\bar{q}'$ and the other $W \to l\nu_l$) is characterized by two quarks coming from the direct decay of top quarks, two quarks coming from the decay of one W and a lepton and a neutrino from the other W decay. Therefore the final experimental signature is four or more jets, a single lepton (electron or muon) and missing transverse energy. The generation of Monte

Carlo signal and background samples is described in [8]. The following results refer to an integrated luminosity of 1 fb^{-1}.

3.1. Selection and Event Reconstruction

The selection starts with the High Level Trigger (HLT) requests: a lepton with enough large p_{T} ($p_{\mathrm{T}} >$ 15 GeV for muons or $p_{\mathrm{T}} > 18$ GeV for electrons). The details of the physics objects reconstruction are in [8] and references therein. Offline electron reconstruction and identification is performed by using tracker and electromagnetic calorimeter information and the muon reconstruction uses both tracker and muon chambers sub-detectors information. An isolation variable for the leptons is defined as the ratio between the sum of p_{T} of the tracks and energies of calorimetric deposits around the candidate and the p_{T} of the lepton candidate itself. The lepton candidate must have such isolation variable less than 0.1 and $p_{\mathrm{T}} > 30$ GeV/c. If more than one lepton is selected, the event is rejected. The jet reconstruction algorithm uses the calorimetric energy deposits with a seed threshold of $E = 1$ GeV and performs an iterative cone procedure with radius $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} =$ 0.5. The jet candidates are selected by requiring $E_{\mathrm{T}} > 40$ GeV and $|\eta| < 2.4$; in order to reject fake jets, they are required to have the fraction of electromagnetic energy to the total energy less than one and to be far enough from the lepton candidate (*lep*) by imposing $\Delta R(jet, lep) > 0.5$. The missing transverse energy ($\not{E}_T$) used in this analysis is computed by performing the vectorial sum of the energy deposits in the calorimeters. The reconstruction of neutrino momentum is needed to compute the leptonic top quark mass; the transverse component comes from the $\not{E}_T$ value while the longitudinal component is determined from the four-momentum conservation of the W boson decay. A useful kinematic variable to reduce the background contamination is Centrality. It represents the fraction of the hard scattering going in the transverse plane and it is defined as:

$$\mathrm{Centrality} = \frac{\sum \mathrm{E_T}}{\sqrt{\left(\sum \mathrm{E}\right)^2 - \left(\sum \mathrm{p_z}\right)^2}} \qquad (2)$$

It is required to be larger than 0.35. The final step of the event reconstruction is the computation of the invariant masses using the selected reconstructed objects. Among the selected jets, the four with largest E_{T} are considered as coming from the decays of the two top quarks and of the hadronic W. While the selected lepton and the missing energy are known to come from a W decay, the assignment of the four chosen jets to the partons has to be determined. In order to choose the right combination, a two step association is used. Beforehand the masses and the widths of the hadronic W boson and the tops are obtained from simulation. The distributions of the three invariant masses of the reconstructed objects well matched to the generated particles are used to obtain the parameters m_{Whad}, m_{tHad}, m_{tLep}, $\sigma(m_{Whad})$, $\sigma(m_{tHad})$ and $\sigma(m_{tLep})$. First the hadronic W boson is reconstructed by computing the invariant mass of every pairs of jets among the four. The pair with the nearest invariant mass to the W one, namely ij, is chosen. The following cut, is applied:

$$|m_{ij} - m_{Whad}| < \sigma(m_{Whad}) \qquad (3)$$

The second step is the association of the two remaining jets (k and p) to the partons coming from the direct decay of top quarks. To this end a χ^2 based on the two top quarks masses is defined:

$$\chi^2 = \left(\frac{m_{ijk} - m_{tHad}}{\sigma(m_{tHad})}\right)^2 + \left(\frac{m_{l\nu p} - m_{tLep}}{\sigma(m_{tLep})}\right)^2 \qquad (4)$$

where i and j are the 2 jets chosen as coming from the W boson decay. Now the only combinatorial ambiguity lies in the choice of which one of the two remaining jets is associated to which of the two top quark. The association that minimizes the χ^2 is assumed to be the correct one. We consider the events with a large χ^2_{min} as events which are wrongly reconstructed, so the cut $\chi^2_{min} < 4$ is applied.

After the whole selection, the expected event number with an integrated luminosity of 1 fb^{-1} is 2650 for semileptonic $t\bar{t}$, while the main background processes are: 109 for other $t\bar{t}$, 260 for $W + jets$, 52 for $Z + jets$, 52 for tW and 56 for QCD di-jet. The b-tagging algorithm adopted in this analysis is the *JetProbability* [10] and the chosen working point is such that $\epsilon_b = (79 \pm 1)\%$ and $\epsilon_q = (13 \pm 1)\%$.

3.2. Background Subtraction

The χ^2_{min} defined in Eq. 4, and referred to as χ^2_{normal} in the following, has a peak at low values of χ^2 for correctly reconstructed semileptonic $t\bar{t}$ events, called *signal* in the following. Background and incorrectly reconstructed $t\bar{t}$ events (*Background* in the following) lead to low values of χ^2_{normal} only due to random combinatorics. Therefore if the direction of one of the selected jets is artificially changed, the mass χ^2 distribution should remain the same for background events, while we expect the distribution for *signal* events will appreciably change. We can define a χ^2_{random} just like the χ^2_{normal}, but computed by assigning a random direction to one of the two jets considered as coming directly from the tops. We decided to change the direction of the one with highest transverse energy. Uniform distributions for ϕ and η have been generated, allowing for ϕ in the range $(-\pi, \pi)$ and η in the range $(-2.4, 2.4)$, as for the selected true

jets. Then the χ^2 procedure was repeated leading up to the new combination that gives the minimum χ^2, called χ^2_{random}. Fig. 1 shows the distribution of the χ^2_{min} variable separately for *signal* and background events. The n_{btag} distribution of the events selected

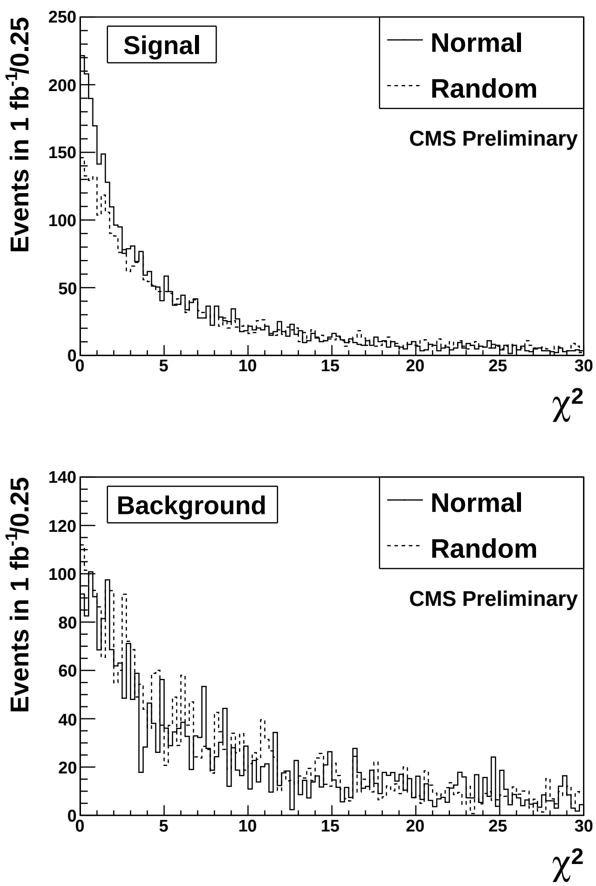

Figure 1: Up: χ^2_{min} distribution of the *signal* sample (as defined in the text). Down: χ^2_{min} distribution of the complete background sample. Both the figures show the χ^2_{normal} (solid) and the χ^2_{random} (dashed) distributions.

after the cut $\chi^2_{normal} < 4$ will be referred as n^{normal}_{btag} while the events selected after the cut $\chi^2_{random} < 4$ will be referred as n^{random}_{btag}; Fig.2 (Upper panel) shows the result of the n^{normal}_{btag}-n^{random}_{btag} subtraction for the *signal* sample (solid) and for the background sample (dashed), the latter is compatible with a flat zero distribution. Therefore, it is clear that if one considers the whole data sample, containing signal and background events, and computes bin-by-bin the difference of the $Normal$ and $Random$ distributions, the resulting n_{btag} distribution will be proportional to the distribution of the *signal* only, as Fig. 2 (down) shows.

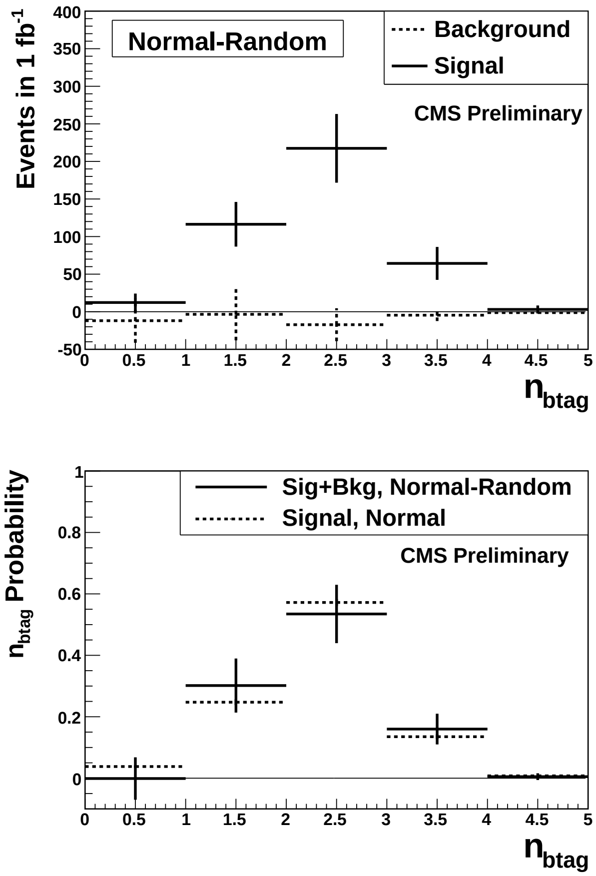

Figure 2: Up: n^{normal}_{btag}-n^{random}_{btag} distribution for *signal* (solid) and *background* (dashed) events normalized to $L = 1$ fb^{-1} . Down: n^{normal}_{btag}-n^{random}_{btag} distribution for the whole data sample (solid) and n^{normal}_{btag} distribution of the only *signal* (dashed) normalized to unity.

3.3. Fit Results

The distribution resulting from the bin-by-bin subtraction of the whole data sample, after normalization, is to be fitted with Eq. 1. In order to check the effectiveness of the method the fit was repeated assuming several R values. Different values of R (R_{gen}) were generated in the range [0.9, 1] by properly weighting three samples where the decay of $t\bar{t}$ was forced: $t\bar{t} \rightarrow WbWb$, $t\bar{t} \rightarrow WbWq$, $t\bar{t} \rightarrow WqWq$. The statistical uncertainty remains steady in all the range and it is $\sigma_R(stat) = 0.12$. The measured values of R agree within the statistical uncertainty with R_{gen} in the range $R_{gen} = [0.9, 1]$.

3.4. Systematic Uncertainties

The various uncertainties were estimated based on the anticipated knowledge of the CMS experiment after 1 fb^{-1} of integrated luminosity [8]. All system-

atic contributions were assumed to be uncorrelated, therefore the total systematic uncertainty has been computed by square summing. In order to check the impact of ϵ_b on R measurement, its value was varied by 4%. Since the number of b-tagged jets from Eq. 1 does not take into account the presence of b-jets from radiation, while the value of ϵ_b measured in real data [12] does, a contribution due to such bias is considered. The systematic uncertainty associated to the jet energy scale is estimated by shifting the calibrated transverse energy for each jet used in the analysis by a relative 5%. The effect of the difference in the selection efficiency between $t\bar{t} \rightarrow WbWb$ (ε_{bb}) and $t\bar{t} \rightarrow WqWq$ (ε_{qq}) was estimated by varying ϵ_{bb} by $\varepsilon_{bb} - \varepsilon_{qq}$ (=0.04%). The χ^2 cut was varied by ± 0.5. The systematics study results for each source are summarized in Table I together with the total value.

Table I Contributions to systematic uncertainty.

systematics	σ_{sys}
b tagging efficiency	0.04
b tagging efficiency bias	0.04
Jet Energy Scale	0.09
χ^2 cut	0.02
Selection efficiency	0.006
total	0.11

4. Dileptonic $t\bar{t}$ analysis

This study considers $t\bar{t}$ events were both the W decay to leptons, the final state with an electron and a muon was chosen, as it is the channel with the largest cross section and smallest background. The generation of Monte Carlo signal and background samples is described in [9]. All the results refer to an integrated luminosity of 250pb^{-1}.

4.1. Event Selection

The event selection is tuned to identify leptonic final states with two prompt, isolated leptons with high transverse momenta in the CMS detector. The selection is detailed in [9]. Data samples are triggered by requiring a non-isolated single muon ($p_T > 9$ GeV/c) or a single electron ($E_T > 15$ GeV). Lepton candidates are reconstructed with $p_T \geq 20$ GeV/c in the fiducial region $|\eta| \leq 2.4$ of the detector and must satisfy identification and isolation requirements. The leptons are required to be separated by $\Delta R > 0.1$. In the case of multiple selected leptons, the ambiguity is resolved by selecting $e\mu$ candidates with opposite electric charge and highest transverse momenta. Jets are reconstructed using the seedless infrared-safe cone algorithm and are required to have at least two calorimeter towers with a minimum E_T sum of 2 GeV. Jets are required to have at least one assigned track so that the b-tagging algorithms can be applied. These cuts define the “taggability” requirements. The energy of the jets is corrected for the η-dependence and absolute E_T using MC based corrections for generator level jets. Taggable jets are selected with E_T(corrected)$\geq$ 30 GeV/c and $|\eta| \leq 2.4$. Jet candidates are further required to be separated from the selected leptons by ΔR(jet, lepton) ≥ 0.3 and to have an electromagnetic fraction EMF<0.98. The total missing transverse energy, $\not{E}_T$, is corrected for the energy deposited by muons and it is required to be above 30 GeV. With 250 pb^{-1} of integrated luminosity, after the described selection, the expected event number is 787 for dileptonic $t\bar{t}$, and the main background contributions are due to other $t\bar{t}$ (14 events),single top (29 events), Di-boson (10.5 events), $W/Z + jets$ (26 events). Therefore after the selection a signal to backround ratio of approximately 10 is expected.

4.2. The jet misassignment estimate

Despite small contributions from other background processes there is a non-negligible probability that at least one jet from a $t\bar{t}$ decay is either missed because it was not reconstructed or because it did not pass the jet selection criteria, and another jet is chosen instead (such as, for example, jets from ISR/FSR). This will be referred to as “jet misassignment” and an estimate of the jet misassignment level has to be made from data. The estimate is done in terms of probability weights α_i, where $i = 0, 1, 2$ is the number of jets from top decays correctly reconstructed and selected. The selected events are a combination of three different categories:

- events with no jet selected from the top decays, weighted by α_0 (background-dominated);
- events with only one jet correctly assigned to the top decay, weighted by α_1 (combination of signal and background);
- events with two jets correctly assigned to the top decays, weighted α_2 (signal-dominated) .

In first approximation the weights α_i can be parameterized in terms of a binomial combination of α, the probability of correctly assigning individual jets. The value of α can be estimated using the kinematic properties of the events directly from data. A correlation can be sought in the lepton-jet pairs originating from the same top quark decay [13] and it is possible to show that no pair with $M_{l,b} > M_{l,b}^{max} \equiv$

$\sqrt{m_t^2 - m_W^2} = 156$ GeV/c^2 should be observed (spectrum endpoint). Two methods are proposed to emulate the invariant mass distribution of the misassigned jets: "swapping" the jet in the assigned lepton-jet pair, with a jet from a different event, or "randomly rotating" the momentum vector of the selected leptons. As the "random rotation" and "swap" methods yield similar results, the average value is used to model the invariant mass distribution of the background jets. The distribution of the "swapped" and "randomly rotated" pairs, normalized to fit the high-end part of the distribution, is superimposed. The two background models provide a good estimate of the fraction of misassigned pairs with $M_{lepton,jet} > 190$ GeV/c^2 (Fig. 3). The normalization factor applied to the distribution of the swapped (randomly rotated) pairs is related to the misassignment fraction, $1 - \alpha$ [9].

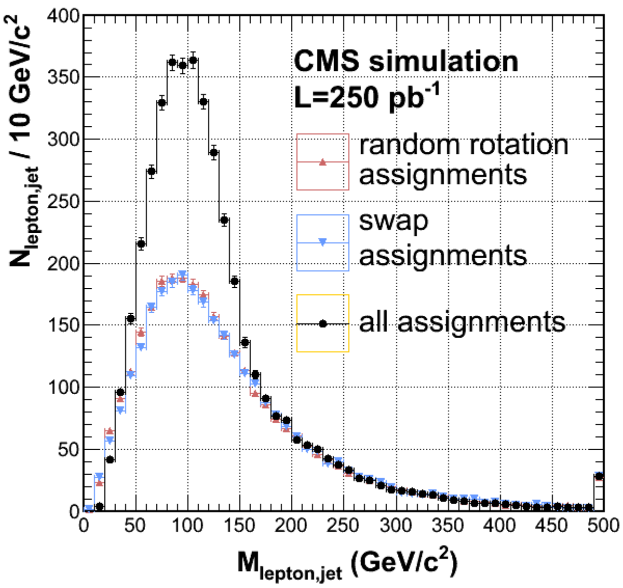

Figure 3: Invariant mass of the lepton-jet pairs for the whole data sample.

4.3. Measurements by fitting the n_{btag} distribution

The following subsections describe the measurement of R and of the b-tagging efficiency respectively; for both the measurements a fit of the n_{btag} distribution is performed with the function in Eq.1 as in the semi-leptonic analysis, but here it depends also on α (besides R, ϵ_b, ϵ_q). In both the studies the $\alpha_2 = \alpha^2$ parameter is fixed to the value obtained by data as explained above, α_0 is a free parameter and α_1 is obtained from the normalization (α_1=1-α_0-α_2). The value obtained for α_2 is $\alpha_2 = 0.67 \pm 0.07$ (stat)$\pm$0.03 (syst), while the one obtained by using the MC truth is 0.63$\pm$0.02. ϵ_q is fixed to the value obtained by other data driven methods [14] and the other parameters are fixed or free depending on the study.

4.3.1. Measurement of R

In order to measure R, ϵ_b was fixed; the results for $R_{gen} = 1$, for the two b-tagging algorithms (JP and TC), each for three working points, are shown in Tab.II Figure 4 shows the results obtained by fitting

Table II R fit results for an integrated luminosity of $L = 250$ pb^{-1}. Statistical uncertainties from the fit and from MC truth are included.

b-tagging algorithm	Working point		
	loose	medium	tight
Jet Probability	1.01 $\pm$ 0.02	1.00 $\pm$ 0.02	0.97 $\pm$ 0.03
Track Counting	1.00 $\pm$ 0.02	0.99 $\pm$ 0.02	1.04 $\pm$ 0.03

R and α_0 using jets tagged with the JP loose point. Different subsamples with forced decays, are weighted

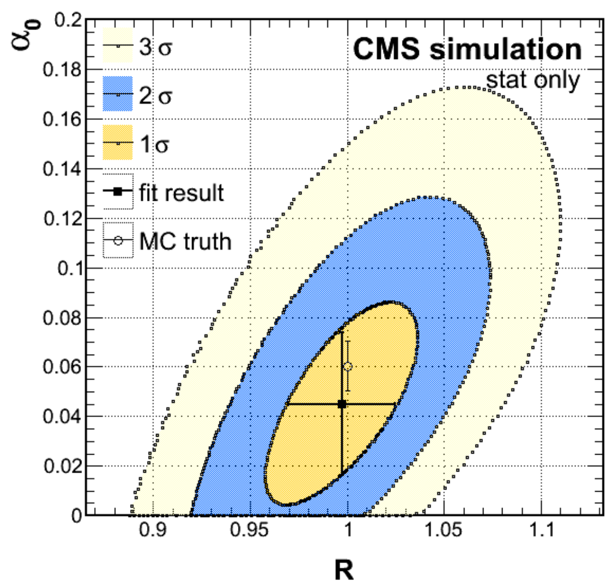

Figure 4: Fit to R and α_0, only statistical uncertainties are shown.

similarly to the semi-leptonic study (Sec.3.3) in order to give different R_{gen} values. All backgrounds are included. The b-tag multiplicity distributions obtained this way are sampled according to the expected number of events and fit to determine R. The statistical uncertainty of each fit result is then determined from the width of the distribution of $R_{generated} - R_{measured}$. The systematic uncertainties are dominated by the uncertainty on the b-tagging efficiency. The total uncertainty is $\sigma_R(stat + sys) = 0.09$ with 250 pb^{-1}.

4.3.2. Measurement of b-tagging efficiency

Here $R = 1$ is fixed and the b-tagging efficiency, ε_b is measured. The constraint $0 \leq \varepsilon_b \leq 1$ is used in the fit. The results are shown in Tab. III. A simultaneous fit to the b-tagging efficiency and α_0 yields the 2-dimensional distribution shown in Figure 5. The total systematic uncertainty is 4% and is due to the uncertainty on α. The uncertainty is estimated by repeating the fit procedure after displacing each parameter by positive/negative values from a Gaussian distribution centered at zero with a width given by the corresponding uncertainty of the parameter. The sensitivity of the ε_b measurement is about ± 0.02 when R is varied by 5%. The fitting model is derived for $t\bar{t}$ events and the bias is estimated to be small, given that the background events are only 10% of the total sample. The good agreement (within uncertainties) of the fit results with the MC truth values justifies this assumption. The uncertainty due to different ISR/FSR content in the final sample is expected to be small ($< 1\%$).

Table III Fit to b-tagging. $R = 1$ fixed and α is fixed to the value estimated with the swap method. Statistical uncertainties from the fit and from MC truth are included.

algorithm	working point	ε_b (MC truth)	ε_b
Jet Probability	loose	0.82 ± 0.01	0.81 ± 0.02
	medium	0.63 ± 0.01	0.63 ± 0.02
	tight	0.41 ± 0.01	0.41 ± 0.02
Track Counting	loose	0.80 ± 0.01	0.82 ± 0.02
	medium	0.65 ± 0.01	0.65 ± 0.02
	tight	0.40 ± 0.01	0.41 ± 0.02

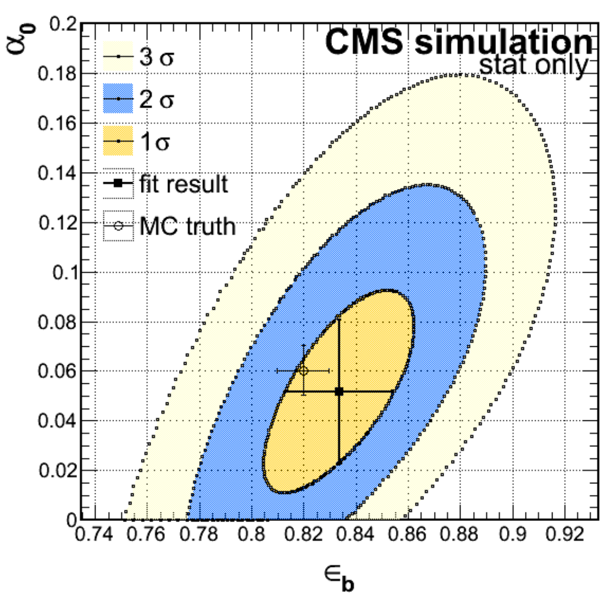

Figure 5: Contour plot of the fit to b-tagging efficiency and α_0. $R = 1$ fixed and α is fixed to the value estimated with the swap method.

5. Conclusions

Two studies of feasibility of the R measurement was presented, one by using selected semi-leptonic $t\bar{t}$ events and the other by using selected di-leptonic $t\bar{t}$ events in the $e\mu$ channel. The expected uncertainties, for the semi-leptonic channel with $L = 1$ fb^{-1}, are $\sigma_R(stat) = 0.12$ and $\sigma_R(sys) = 0.11$. For the di-leptonic channel, with $L = 250$ pb^{-1}, the expected uncertainty is $\sigma_R(stat + sys) = 0.09$. Furthermore, in the dileponic channel a study on the ϵ_b measurement, fixing R to the SM value, has been performed. The expected uncertainties are: $\sigma_{\epsilon_b}(stat) = 0.02$ $\sigma_{\epsilon_b}(sys) \sim 0.04$. Both the studies use data driven methods to subtract the background contribution.

References

[1] J. Awall et al *Eur. Phys. J.* C49 791-801(2007).
[2] D. Acosta et al., CDF Collaboration, *Phys. Rev. Lett.* 95 102002(2005)[hep-ex/0505091].
[3] V. M. Abazov et al., D0 Collaboration, *Phys. Rev. Lett.* 100,192003(2008).
[4] C. Amsler et al., *Phys. Lett.* B667,1(2008).
[5] V. M. Abazov et al. [The D0 Collaboration], arXiv:0903.0850[hep-ex].
[6] T. Aaltonen et al. [The CDF collaboration], arXiv:0903.0885 [hep-ex].
[7] CMS Collaboration, *The CMS experiment at the CERN LHC* JINST 3:S08004,2008.
[8] CMS Physics Analysis Summary TOP-09-007 http://cdsweb.cern.ch/record/1194522?ln=en
[9] CMS Physics Analysis Summary TOP-09-001 http://cdsweb.cern.ch
[10] *CMS Physics Technical Design Report Vol.1*,CERN/LHCC 2006-001.
[11] CMS PAS BTV-07-001.
[12] CERN-CMS-NOTE-2006-019.
[13] Ellis, R.K. et al., Cambridge Monographs on Part. Phys., Nucl. Phys. and Cosmology.
[14] CMS PAS BTV-07-002.